Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 119-126 recibido 29-04-2005; aceptado 01-06-2005

Text Categorization using bibliographic records: beyond document content

Arturo Montejo-Ráez
Dep. of Computer Science
Universidad de Jaén
Jaén, Spain
amontejo@ujaen.es

L. Alfonso Ureña-López
Dep. of Computer Science
Universidad de Jaén
Jaén, Spain
laurena@ujaen.es

Ralf Steinberger
IPSC
Joint Research Center
European Commission
Ispra, Italy
ralf.steinberger@jrc.it

Resumen: En este artículo se estudia el uso de diferentes fuentes de información para tareas de clasificación de textos. Dado el creciente número de bibliotecas digitales, se impone una revisión de la información disponible en dichas bases de datos. Se han llevado a cabo una serie de experimentos de clasificación multi-etiquetado dentro del dominio de la Física de Altas Energías haciendo uso de diferentes clasificadores base y combinando distintas fuentes de información. Los resultados muestran que el uso de metadatos es tan válido como el uso de versiones a texto completo de los documentos.
Palabras clave: clasificación automática de documentos, aprendizaje automático, bibliotecas digitales.

Abstract: This paper studies the use of different sources of information for performing a text classification task. The growing number of digital libraries imposes a review of the available data from those databases. Some experiments applying different base classifiers for a multi-label classifier in the domain of High Energy Physics on several of these possible sources have been carried out. Results show that the use of metadata is almost as good as the full-text version of papers.
Keywords: text categorization, machine learning, digital libraries.

1 Introduction

Corpora available for Text Categorization (Sebastiani, 2002) research tend to be not as accessible and complete as the research community wishes. Well known collections as *Reuters-21578*[1], *OHSUMED* (Hersh et al., 1994) (used in TREC evaluation forum) or *20 Newsgroups*[2] show, for each sample, sort fragments of text. Instead, within EUROVOC related experiments ((Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat, 2003),(Ralf Steinberger, Johan Hagman, and Stefan Scheer, 2000) and (Ralf Steinberger, Bruno Pouliquen, and Johan Hagman, 2002)) full-text documents are used as sample data for classifiers training. But, in most of the experiments arranged based on these collections, just plain text data from main content of the document is used. Words are then stemmed or lemmatized, counted and weighted following the same indexing scheme.

We have organized this paper by first providing a introduction to bibliographic metadata in CERN databases. Then, the design of the multi-label classifier is described specifying all the techniques involved. Since we want to prove that metadata sources can represent a very enriched source of information improving classification tasks, we have tested them against different classification schemes in order to validate the excelence of these type of sources without depending on specific learning algorithms. Later, the results obtained are listed using macroaveraging over documents for some of the best known evaluation measures. Finally, these results are dicussed and conclusions are reported.

2 Metadata

Data from digital libraries is much richer than just full-text versions of documents: digital libraries contain *metadata*, i.e. additional information about every stored document. Metadata is *data about data*: author of a document, date of publication, storing

[1] Prepared by David D. Lewis. The collection is freely available from the web page `http://www.research.att.com/~lewis/reuters21578.html`

[2] Available at `http://kdd.ics.uci.edu/databases/20newsgroups/20_newsgroups.tar.gz`

format, identifier within the database, publisher, length and so on. The *Dublin Core Metadata Initiative*[3] (DCMI) is a open standard for adding information to documents in digital libraries. The *Open Archive Initiative*[4] (OAI) aims to stablish a standard for document exchange between different digital libraries and a protocol for harvesting and retrieving of documents from database supporting it. OAI also support MARC[5] format for metadata. DCMI is also used as a source of entities for the *Semantic Web*[6] project (Berners-Lee, Hendler, and Lassila, 2001). As we can see, metadata is something to care about.

For High Energy Physics (HEP) related papers stored at CERN (Dallman and Le Meur, 1999; Montejo-Ráez and Dallman, 2001), the European Laboratory for Particle Physics (in Geneva, Switzerland), the MARC format is used (although full accessing through OAI has been recently integrated). A document record sample belonging to this database (CERN Document Server[7]) is shown in figure 1. We have prepared a corpus for our experiments that consists on what we call the *hep-ex* partition, since it is only about experimental Particle Physics. This partition contains 2967 documents and 2793 main keywords, and can be obtained by contacting the authors. These keywords come from human assignment by DESY experts by using the DESY HEPI thesaurus. The DESY Documentation group in Hamburg developed the HEPI (High Energy Physics Information) system from 1963 onwards. In this scheme all documents in the HEP field are indexed by subject specialists who read the entire articles. The thesaurus used contains approximately 2500 terms and has in general been updated every 1-2 years. The terms in this keyword list are used by the DESY Documentation Service for the indexing of papers on high energy (beam momentum above 400 MeV (per nucleon)) and particle physics, accelerator and detector technology and quantum field theory. Though the thesaurus proposes a two-level indexing, only labels on first level have been considered in our experiments, as an approach to reduce the high class imbalance reported by (Montejo-Ráez, Steinberger, and Ureña-López, 2004)). Manually assigned labels to sample document shown in figure 1 are listed in figure 2.

```
<?xml version="1.0" encoding="UTF-8"?>
<collection>

 <dc xmlns="http://purl.org/dc/elements/1.1/"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://purl.org/dc/elements/1.1/
 http://www.openarchives.org/OAI/1.1/dc.xsd">

  <language>eng</language>

  <creator>Albrecht, H</creator>

  <title>
   Determination of the Michel Parameters $\rho$, $\xi$,
   and $\delta$ in $\tau$-Lepton Decays with $\tau \to
   \rho \nu$ Tags
  </title>

  <subject>
   Particle Physics - Experimental Results
  </subject>

  <identifier>
   http://preprints.cern.ch/ ... id=9711022
  </identifier>

  <description>Using the ARGUS detector at the $e^+
  e^-$ storage ring DORIS II, we have measured the
  Michel parameters $\rho$, $\xi$, and $\xi\delta$
  for center of mass energies in the region of the
  $\Upsilon$ resonances. Using 0.04 \pm 0.08$,
  $\xi_{e}= 1.12 \pm 0.20 \pm 0.09$, $\xi\delta_{e}
  = 0.57 \pm 0.14 \pm 0.07$, $\rho_{\mu}= 0.69 \pm
  0.06 \pm 0.08$, $\xi_{\mu}= 1.25 \pm 0.27 the
  combined ARGUS results on $\rho$, $\xi$, and
  $\xi\delta$ using this work on previous
  measurements.</description>

  <date>1997-12-01</date>

 </dc>

</collection>
```

Figure 1: Sample for metadata information in XML DC format.

```
DESY HERA Stor
electron positron
experimental results
magnetic detector
Michel parameter
tau
```

Figure 2: Main DESY keywords manually assigned to sample at figure 1

We can see in those figure 1 the amount of additional data available. The target of this experiment is to test whether the use of additional data (appart from content data)

[3] http://dublincore.org/
[4] http://www.openarchives.org
[5] http://www.loc.gov/marc/
[6] http://www.w3.org/2001/sw/
[7] http://cds.cern.ch

can improve classification performances. Due to the profusion of meta-data entries in the records of current digital libraries, a positive answer is an important clue to build enriched classifiers.

3 Multi-label classifier

TECAT stands for TExt CATegorization. It is a tool developed by authors for building multi-label automatic text classifiers, assigning more than one class per document (Montejo-Ráez, Steinberger, and Ureña-López, 2004). With TECAT you can experiment with different collections and classifiers in order to train an automatic multi-labeled text classifier.

TECAT implements the whole training and classification process for a multi-label classifier based on machine learning algorithms. A better look of the TECAT way-of-work can be given by describing every phase involved in training and classification along with its related parametrization.

1. **Indexing**. This phase corresponds to the task of selecting and filtering features (terms) that will represent each document. No multiwords identification has been considered:
   - Stop words have been removed.
   - Words have been truncated by passing them to the Porter's Stemming algorithm (Porter, 1997). From now onwards we refer to these resulting items as *terms* or *features*.
   - Terms with less than 40 characters have been discarded, to avoid malformed long words (result of a bad output from the PDF converter). This number has been set empirically.
2. **Folding**. This stage performs a stratified 10-fold partitioning of the collection (Mitchell, 1997; Kohavi, 1995). At this stage additional filtering and feature selection is performed:
   - The 50.000 terms with the highest information gain value have been kept. This limit has been chosen empirically from paralel experimentation.
   - The classical `TF.IDF` $term - frequency \cdot inverse - document - frequency$ (Salton and Buckley, 1988) weighting scheme has been used to weight features.
   - Term weights in document vectors have been normalized using the cosine normalization (to force all vectors having the same norm).
3. **Learning**. This stage corresponds to learning algorithms (Lewis et al., 1996).
   - The base algorithms used have been Rocchio, Widrow-Hoff (Lewis et al., 1996), PLAUM (Y. et al., 2002; S., S., and H., 2001) and Support Vector Machines (Joachims, 1998).
   - A threshold has been computed according to the S-cut approach (Yang, 2001) and applied to Rocchio and Widrow-Hoff algorithms.
   - The minimal value that the *F1* measure (van Rijsbergen, 1975; Lewis, 1991) must reach in order to consider the classifier as valid candidate (Montejo-Ráez, Steinberger, and Ureña-López, 2004) has been set to 0.1. With this approach, classes for which classifier are not enough performance are just discarded (we will just do not predict them).

For each of these algorithms, a different corpus of data has been used for learning and testing. Therefore, several corpora are produced as a combination of full-text information with available metadata:

- **Source A: Abstracts**. Experiments with given algorithms have been carried out on abstracts (the *description* field in XML DC form).
- **Source M: Metadata**. Each document is composed by a combination of content data (*abstract* and *title*) with fixed values (*date*, *subject*, *creator* and *language*).
- **Source F: Full-text**. The corpus is composed by the full-text version of documents (extracted from PDF versions of each paper).
- **Source F+M: Full-text and metadata**. This is the most complete corpus, where plain-text data is built up from full-text version, title and abstract, and

combined with fixed fields like *date*, *subject*, *creator* and *language*.

It is important to note that the processment of content data differs from that applied on metadata fields. These fields have been protected to not be affected by stop words removal and stemming procedures. For the rest of operations, each field is considered and additional feature.

```
1.233175 tau
1.533310 electron positron
0.327726 magnetic detector
```

Figure 3: Main DESY keywords automatically assigned by TECAT to sample at figure 1. They are preceded by their associated *Classification Status Value* (in this sample, returned by the PLAUM algorithm).

Three measures have been studied to determine whether a source outperforms another. These measures are *precision*, *recall* and *F1*. F1 provides a more general view than precision and recall separatedely, being a common measure of the goodness of the method (though not very preferred in text processing tasks). However, precision and recall have been studied appart to inspect how a source affects them.

A sample of the output produced by TECAT is shown in figure 3. These labels have been selected automatically for sample given in figure 1. As we can see, just three keywords are selected by the classifier, with 100% of precision and 50% of recall, though.

4 Description of experiments and results

The description of experiments is described as follows: for each source (corpus), we have run four different experiments. Each of such experiments consists in using a different unique base classifier of the four mentioned before. As results, we will have sixteen sets of results. The reason to use different algorithms is to not be so dependent on the type of learning method used. Therefore, results for the four algorithms will be considered together and compared to those for experiments based on a different corpus. At each run the three measurements (precision, recall and F1) have been obtained for the ten most frequent keywords (see table 1).

Then, a Wilcoxon Signed Raked test has been carried out. This is a non-parametric

```
experimental results
magnetic detector
talk
electron positron
quark
CERN LEP Stor
anti-p p
Z0
Batavia TEVATRON Coll
mass spectrum
```

Table 1: Ten most frequent categories

test that let us know if two distributions on the same variable are statistically different. We have used *Octave*[8] to compute the *p-values* returned by the test. In our comparison, we will consider only p-values under 0.05, that is, the probability for the two given distributions to differ not by chance is higher than 95%. Since we have run a 10-fold cross-validation mechanism within TECAT, we have averaged the measured values of precision, recall, F1 and accuracy for the ten most frequent keywords. Therefore, to compare source A (abstracts) against source F (full-text), we created two distributions with 40 values each (4 algorithms $\times$ 10 measured keywords). Then, we could construct the comparison matrices (one per measure) with Wilcoxon p-values shown at tables 2, 3 and 4.

By looking at those tables we can determine the confidence that two source strategies are different. For example, as seen on table 2, the Wilcoxon test found a p-value of 0.000045 between strategies **F+M** and **A**. We can interpretate this value as a confidence of 99.9965% that the two distributions are different (based on precision measurements in this case). Besides, because of the tailored nature of our test, this measure also indicates the confidence on asserting that strategies in colums are *better* than strategies in rows.

on prec.	**A**	**M**	**F**	**F+M**
A	0.5000	0.0627	**0.00014**	**0.000045**
M	0.9372	0.5000	0.0790	**0.026**
F	0.9998	0.9209	0.5000	0.09387790
F+M	0.9999	0.9735	0.9061	0.5000

Table 2: Tailored Wilcoxon test over precision

[8]Octave is a high-level language, primarily intended for numerical computation that is available under the GPL. `http://www.octave.org`

on rec.	**A**	**M**	**F**	**F+M**
A	0.5000	**0.00000002**	**0.0018**	**0.000001**
M	0.9999	0.5000	0.9997	0.7318
F	0.9981	**0.00023**	0.5000	**0.000001**
F+M	0.9999	0.2681	0.9999	0.5000

Table 3: Tailored Wilcoxon test over recall

on F1	**A**	**M**	**F**	**F+M**
A	0.5	**0.0000001**	**0.000018**	**0.00000008**
M	0.9	0.5000	0.9766	0.2217
F	0.9	**0.0233**	0.5000	**0.00000007**
F+M	0.9	0.7782	0.9999	0.5000

Table 4: Tailored Wilcoxon test over F1

Wilcoxon can be one-tailored or two-tailored. For one-tailored test we can conclude if a difference exists, but we cannot say anything about the *direction* of it. It is our interested to find if a certain source of data provides better results *over* another source of data. Thus, we are interested in a one-tailored analysis that will let us know if one source outperforms another source. This is how it has been computed in the given tables.

The macroaveraged values by document obtained in our experiments are given in table 5. At this table, each row shows the measures of performance for a given algorithm by using a given source. For example, the entry *WH M* specifies measured macroaveraged values by document using the metadata based corpus as source and applying Widrow-Hoff algorithm as base classifier (with S-cut thresholding as detailed previously in configuration description).

Many classes are just discarded (column *% classes*)in the filtering process or due to fact that no possible algorithm can be trained with a minimum performance (the percentage column provides useful information about it). It is interesting to note how different algorithms are more "trainable" on some classes than others at this point. Anyhow, performance measures are obtained considering all classes covered by the corpus, not only trained ones. It is significative how some algorithms (PLAUM and SVM) which a reduced set of candidate classes report as good performance as other algorithms covering a wider range of classes (Widrow-Hoff and Rocchio).

5 Analysis of results

First, we remind the target of these experiments: to study how the use of different sources of data affects performance of a multi-label categorization engine. Not only is interesting to find which combinations of sources are best, but also to study whether is worth using one over another. For example, the computational cost of using full-text documents is much higher than that of using just small abstracts: the collection is smaller, the trained data demands lower storage space, and the classification process is sped up. Thus, it is very important, in our domain, to identify such facts.

Prec.	**Recall**	**F1**	**% classes**	**Experiment**
0,4429	0,5403	0,4553	84,43	WH A
0,4412	0,5555	0,4641	85,00	WH F
0,4553	0,5534	0,4663	84,54	WH M
0,4639	0,5713	0,4805	87,27	WH F+M
0,4711	0,5421	0,4611	88,16	Rocchio A
0,4272	0,5235	0,4281	86,12	Rocchio F
0,4525	0,5605	0,4569	86,18	Rocchio M
0,4424	0,5419	0,4432	87,92	Rocchio F+M
0,6911	0,4104	0,4895	52,48	PLAUM A
0,7106	0,4348	0,5118	57,67	PLAUM F
0,7201	0,4486	0,5266	55,67	PLAUM M
0,7254	0,4529	0,5314	59,27	PLAUM F+M
0,7458	0,3336	0,4342	31,14	SVM A
0,7549	0,3574	0,4590	35,03	SVM F
0,7734	0,3516	0,4586	32,23	SVM M
0,7697	0,3732	0,4775	36,62	SVM F+M

Table 5: Macroaveraged measures for all classes in performed experiments

To understand these results we can notice at table 4, for instance, that the p-value corresponding to row **F** and column **F+M** is very small: 0.00000007. It says that when using the corpus created as combination of full-text and metadata, we can be sure (99.999993% sure!) of obtaining more accurated results than those obtained from corpus based on just full-text data. We also find that the p-value shows a *transitive* behaviour: let a, b and c, be distinctive sources, and '$>$' a binary operator to indicate that a source provides better results than another, then, we can state that:

$$(a \geq b) \wedge (b \geq c) \Longrightarrow a \geq c \qquad (1)$$

At table 4 **M** outperforms **A**, and **F+M** outperforms **M**, therefore, we can see that **F+M** outperforms widely **A**.

It is possible to illustrate graphically the improvement obtained when using some sources instead of others. Figures 4, 5 and 6 show the effect, per measure, of each source. At each diagram the considered

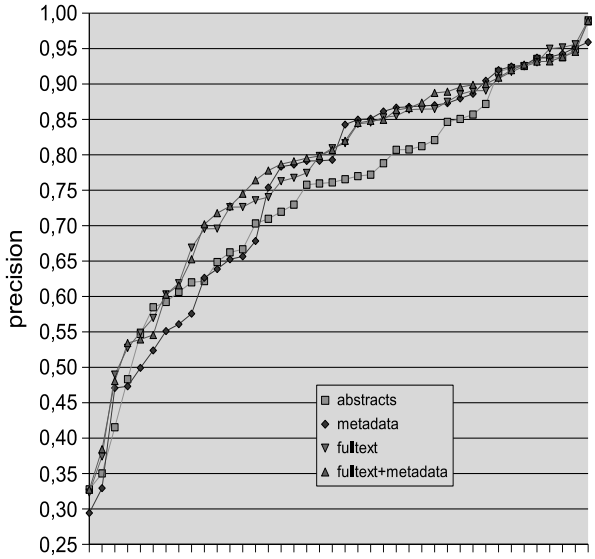

Figure 4: Comparison of sorted measures for each source over precision

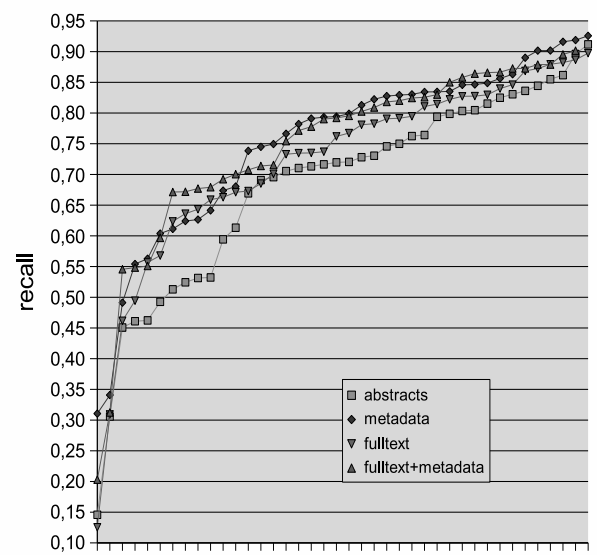

Figure 5: Comparison of sorted measures for each source over recall

sources are drawn, with their 40 measured values sorted to ease the visual comparison between sources. Consequently, each source series is composed by 40 points. In principle, the use of only the abstract of a document seems to be the worst choice. Actually, this is not an easy question, as we will argue from Wilconox test results.

As said before, we can notice an increment in performance just having a look at those diagrams, but the statistical values contained in tables 2, 3 and 4 provide solid evidences on the studied subject.

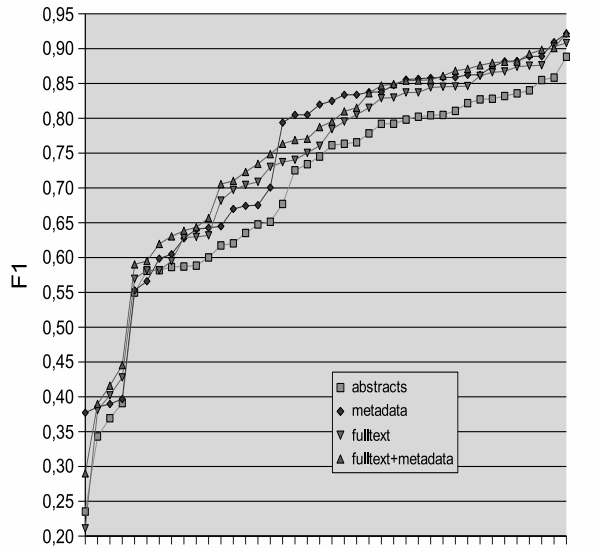

Figure 6: Comparison of sorted measures for each source over F1

6 Conclusions

By reviewing one table after the other we can conclude the following points:

1. The combined use of metadata and full-text information is the best choice in any case (though the most costly).
2. Metadata is also good choice in most cases despite the fact that is outperformed by the combination of full-text and metadata when we are more interested in precision.
3. Metadata source even outperforms corpus based on full-text papers for recall and F1.
4. As main conclusion, **the metadata source should be preferred** due to its reduced computational cost and its good behaviour against full-text and (except for precision matters) combined full-text and metadata.

Therefore, adding a big full-text content of a document (extracting it from the PDF version) to the features that can be promptly extracted from the document record stored in the database is not worthy. This conclusion validates the visual conclusion obtained from inspecting previous diagrams, where we can see how the curve for the metadata source is usually above the rest of curves. This effect is less clear for precision, as reported by the test.

7 Acknowlegdements

This work is partially financed by the Spanish Minister of Science and Technology, by means of project TIC2003- 07158-C04-04.

References

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American*, 284(5):34–43, May.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. 2003. Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In Amalia Todirascu, editor, *Proceedings of the workshop 'Ontologies and Information Extraction' at the EuroLan Summer School 'The Semantic Web and Language Technology'(EUROLAN'2003)*, page 8 pages, Bucharest (Romania).

Dallman, David and Jean-Yves Le Meur. 1999. Automatic keywording of high energy physics. Technical report, European Laboratory for Particle Physics, Geneva, Switzerland, October.

Hersh, William, Chris Buckley, T. J. Leone, and David Hickam. 1994. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201. Springer-Verlag New York, Inc.

Joachims, Thorsten. 1998. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

Kohavi, Ron. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1145. Morgan Kaufmann, San Mateo, CA.

Lewis, D. D. 1991. Evaluating Text Categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 312–318. Morgan Kaufmann.

Lewis, David D., Robert E. Schapire, James P. Callan, and Ron Papka. 1996. Training algorithms for linear text classifiers. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pages 298–306, Zürich, CH. ACM Press, New York, US.

Mitchell, Tom M., 1997. *Machine Learning*, chapter Decision tree learning, pages 55–59. McGraw-Hill, New York.

Montejo-Ráez, A., R. Steinberger, and L. A. Ureña-López. 2004. Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. In Vicedo J. L. et al., editor, *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004*, number 3230 in Lectures notes in artifial intelligence, pages 1–12. Springer.

Montejo-Ráez, Arturo and David Dallman. 2001. Experiences in automatic keywording of particle physics literature. *High Energy Physics Libraries Webzine*, (issue 5), November. URL: http://library.cern.ch/HEPLW/5/papers/3/.

Porter, M. F., 1997. *An algorithm for suffix stripping*, pages 313–316. Morgan Kaufmann Publishers Inc.

Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Third International Conference on Intelligent Text Processing and Computational Linguistics*.

Ralf Steinberger, Johan Hagman, and Stefan Scheer. 2000. Using thesauri for automatic indexing and for the visualisation of multilingual document collections. pages 130–141.

S., Robertson, Walker S., and Zaragoza H. 2001. Microsoft cambridge at trec-10: filtering and web tracks. In *Text Retrieval Conference (TREC-10)*.

Salton, G. and C. Buckley. 1988. Term weigthing approaches in automatic text retrieval. *Information Processing and Managemnet*, 24(5):513–523.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

van Rijsbergen, C. J. 1975. *Information Retrieval*. London: Butterworths. http://www.dcs.gla.ac.uk/Keith/Preface.html.

Y., Li, Zaragoza H., Herbrich R., Shawe-Taylor J., and Kandola J. 2002. The perceptron algorithm with uneven margins. In *Proceedings of the International Conference of Machine Learning (ICML'2002)*.

Yang, Yiming. 2001. A study on thresholding strategies for text categorization. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval*, pages 137–145, New Orleans, US. ACM Press, New York, US. Describes RCut, Scut, etc.